BEBR

FACULTY WORKING
PAPER NO. 822

On the Distributional Implications of the Coase Theorem and the Core

Jonathan A. K. Cave

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois, Urbana-Champaign

FACULTY WORKING PAPER NO. 822

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

November 1981

On the Distributional Implications of the Coase Theorem and the Core

Jonathan A. K. Cave, Assistant Professor
Department of Economics

Abstract

In this paper we formulate a game of private precautions against externalities and obtain a strengthened version of the Coase Theorem: not only is the result of trading efficient in that it minimizes social cost, but the distribution of inframarginal gains is constrained to lie in the core. We show that the core is nonempty, and provide a decentralized rationale for the core deriving from strong perfect equilibria of the supergame. In addition, the effects of discounting on the set of stable contracts (strong supergame equilibria) and credible stable contracts (strong perfect supergame equilibria) are investigated.

On the Distributional Implications of the Coase Theorem and the Core

I. Introduction

This paper has two objectives. The first is to strengthen, in a specific model, the implication of the Coase Theorem that individual parties will reach a social-cost-minimizing level of precautionary activity independently of the liability rule imposed by the courts. We shall provide two justifications for our stronger result, which is that the distribution of social and private costs that result from such private trading lies in the core of an appropriately-formulated game. The second objective is to point out that there are certain pitfalls in the use of core analysis for such situations, and is motivated by the recent appearance of several articles[1,2] that suffer from shortcomings of this sort.

In the original Coase result, it was demonstrated that private parties would bargain with each other until the marginal cost to the party providing each precautionary activity (including the marginal reduction in his own liability, if any) equalled the marginal benefit to the other party(s). Various authors, including Brown[3] and Brown and Holohan[4] have demonstrated this result, and it is fairly well-known [see Sec. II for a proof] that this results in a level of activity which minimizes the sum of private and social costs, providing that the liability rule is a cost-sharing rule. If it is not, as in the case of the Draconian or eye-for-an-eye rule, the outcome is socially efficient only so long as trading is <u>not</u> allowed.

However, the original result predicted little beyond this efficiency result; in particular, the division of the inframarginal gains was left unspecified. In comparing extreme liability rules, such as no-fault or

caveat emptor, we can conclude that each party would prefer to be in the situation where (s)he begins with no liability, but any further characterization of the result depends on the relative bargaining strengths of the various parties. One way to sharpen the predictive power of this model of bargaining is to use the structure of the model to tell us something about the bargaining strengths of the parties. For example, in a model with only two participants, we would not expect that any participant could extract so much from another that the latter was left in a position inferior to that which (s)he could ensure by unilateral precautionary activity. Extended to the case of more than two participants, this type of reasoning gives us an outcome in the core: the set of outcomes with the property that no group can, by independent action, guarantee an outcome better for each of its members.

This sort of thinking is appropriate to a world in which the parties bargain over the terms of an agreement which is to be enforced by an outside agency. It reflects a basic notion of what is likely to be acceptable, but there is no clear stress on what sort of agreement is actually enforceable. This was pointed out by Professor Coase[2] in a recent comment on the use of the core to predict the outcome of bargaining. He proposed that a better approach might be to look at the type of penalty clauses agents could write into contracts, and see what agreements are actually enforceable by the use of such penalties, which require no outside enforcement. We shall demonstrate that this approach leads to precisely the same set of outcomes as the core.

It might also be argued that, in cases where implementation of a penalty clause against one or more parties found to be in breach of a

contract is costly to the remaining parties who must carry out the punishment, the threatened penalty may lose some credibility. In other words, if party A is held to an agreement only by the threat of a very costly penalty to be inflicted by party B, party A may not believe that party B will carry out the punishment in the event. We shall further demonstrate that the requirement that all threats be credible (in the sense that each party always finds it to be in his best interest to carry out any threat) still leaves us with the same set of outcomes.

In this way, we shall have provided both a centralized (or cooperative) and a decentralized (or non-cooperative) justification for predicting an outcome in the core. However, there is one final issue to be dealt with. If it happens that that core is empty, then neither of these justifications has much predictive value. If, for every proposal, there is some group of individuals that can improve upon the proposal by independent action, it is difficult to have confidence in any particular outcome, let alone a socially efficient one. This is the point made by Professors Aivazian and Callen[1] in a recent article: if there are more than two players, the core may be empty. However, the game we are dealing with is of a fairly specific nature, and we shall be able to demonstrate that, under fairly mild assumptions, the core is not empty.

Finally we shall have some remarks directed to the use of this particular model [the core of a game with unrestricted sidepayments and transferrable utility] as a description of situations involving risk and precaution.

II. The Basic Model and Nonemptiness of the Core

This model is adapted from that used by Brown[3] and Brown and Holohan.[4] Interested readers are referred to those articles for further details.

There are n participants, or agents, indexed f = 1,...,n. We shall use N to refer to the set of all agents.

Each agent $f \in N$ may take a level of precautionary activity x_f, chosen from some compact convex set X_f of possible actions. This results in a private cost of $C_f(x_f)$ to be paid by agent f. We assume that the private cost of each agent's action depends only on his own level of precaution, and not on the levels of precaution adopted by the other agents, although this assumption can be relaxed. It is further assumed that C_f is an increasing and convex function.

Remark: X_f may be multi-dimensional, so that several sorts of precautionary activities may be engaged in by the same agent.

If each agent f has chosen a level of precautionary activity x_f, there results a social cost $\bar{C}(x)$, where $x = (x_1,...,x_n)$. $\bar{C}$ is assumed to be convex and decreasing in all its arguments. This is an important assumption since it is vital for the nonemptiness of the core that all externalities be beneficial to other agents. In addition to the above structure, we posit the existence of a liability rule, which is a function $L: X_1 x ... x X_N \to R^N$. Each element $L_f(x)$ of the vector L(x) measures the proportion of the social cost which must be borne by agent f. We can distinguish certain important special classes of liability rules:

If $L_f(x) \geq 0$ for all f and x, then L is said to be non-indemnificatory;

if $L_f(x) \leq 1$ for all f and x, then L is said to be non-penalizing;

if $\sum_{f \in N} L_f(x) = 1$ for all x, then L is said to be cost-sharing; and

if L(x) is independent of x, then L is said to be constant.

Many of the liability rules implemented by courts are of the cost-sharing variety, but not all: the ancient rule of an eye for an eye,

which may be represented by setting $L_f(x) = 1$ for all x and f, is clearly not a cost-sharing rule, neither are its modern counterparts like public enforcement of antitrust and consumer protection laws without retrospective relief.

We can distinguish between outcomes in this situation on the basis of <u>efficiency</u> (the size of the sum of private and social costs) and also on the basis of the mechanism which leads to the outcome.

We shall say that an outcome is <u>efficient</u> if it minimizes the total cost to society, net of penalties and indemnity payments. This total cost is given by $\sum_{f\in N} C_f(x_f) + \bar{C}(x_1,\ldots,x_n)$, so that in circumstances where the C_f, L and $\bar{C}$ are differentiable, we may characterize efficient outcomes by:

$$(1) \qquad \text{for all } f,\ C_f'(x_f) + \frac{\partial \bar{C}(x)}{\partial x_f} = 0.$$

If each agent pursues his own interests, taking as given the actions of the other agents, we can calculate the action of agent f that minimizes his own cost subject to the constraint imposed by the other agents' choices. Denoting the n-1 vector of these other choices by $x_{(f)} = (x_1,\ldots,x_{f-1},x_{f+1},\ldots,x_n)$, we can define agent f's <u>best response correspondence</u> $b_f(x_{(f)})$ by:

$$(2) \qquad b_f(x_{(f)}) = \{x_f:\ \text{for all } x_f',\ C_f(x_f') + L_f(x_{(f)},x_f')\bar{C}(x_{(f)},x_f') \geq C_f(x_f) + L_f(x_{(f)},x_f)\bar{C}(x_{(f)},x_f)\}$$

In other words, $b_f(x_{(f)})$ represents the best actions that f can take, given the actions $x_{(f)}$ of the other agents.

We would anticipate, in such a noncooperative situation, that the system would come to rest at a situation where each agent is responding optimally to the choices of all the others. This is called a Nash Equilibrium; formally, it is a vector x^n with the property that, for each f, $x_f^n \in b_f(x_{(f)}^n)$. In the differentiable case, we can represent this by the condition

$$(3) \qquad \text{for each f, } C_f'(x_f^n) + \frac{\partial L_f(x^n)}{\partial x_f} \cdot \overline{C}(x^n) + L_f(x^n) \cdot \frac{\partial \overline{C}(x^n)}{\partial x_f} = 0.$$

By comparing this equation with equation (1), it can be seen that there is no necessity for this equilibrium to be efficient; in general, it will not be, unless

$$(4) \qquad \text{for each f, } (1 - L_f(x^n))\frac{\partial \overline{C}(x^n)}{\partial x_f} = \frac{\partial L_f(x^n)}{\partial x_f} \overline{C}(x^n)$$

For example, this is clearly true for the eye-for-an-eye rule, so this results in the socially optimal actions.

On the other hand, if we broaden the degree of cooperation to allow trade or tranactions between agents, we can reach another outcome, which we shall call the Coase Equilibrium, and denote by x^c. It is characterized by the condition that the marginal cost of x_f^c equals the sum of the marginal benefits to all the agents:

$$(5) \qquad \text{for each f, } C_f'(x_f^c) + \sum_{g=1}^{N} \left[\frac{\partial L_g(x^c)}{\partial x_f} \overline{C}(x^c) + L_g(x^c)\frac{\partial \overline{C}(x^c)}{\partial x_f}\right] = 0$$

We remark that, if L is a cost-sharing rule, $\sum_{g=1}^{N} L_g(x) = 1$ for all x so that $\sum_{g=1}^{N} \frac{\partial L_g(x)}{\partial x_f} = 0$ for all x and f. Inserting these two conditions in equation (5) gives equation (1), so we obtain the following conclusion:

Proposition 2.1: If L is any differential cost-sharing liability rule, the Coase Equilibrium is efficient.

This is the result that we shall call the Coase theorem: we note also that allowing trading may destroy efficiency as well. In the case of the eye-for-an-eye rule, which had an efficient Nash Equilibrium, allowing trade moves us to a Coase equilibrium where the relevant marginal condition is

$$(6) \qquad \text{for each f, } C_f'(x_f^c) + n\frac{\partial \overline{C}(x^c)}{\partial x_f} = 0.$$

By the assumed properties of C_f and $\overline{C}$, this means that each agent will be led to take more than the socially-optimal level of precaution.

A more serious problem is that the final distribution of wealth at Coase equilibrium is largely arbitrary. The conditions for Coase equilibrium are marginal conditions, and the allocation of the infra-marginal gains is left unspecified. To see what this means, consider a simple example: there are two agents, and the constant liability rule $L_1(x) = 1$, $L_2(x) = 0$. [Depending on the identities of the agents, this is either caveat emptor or caveat venditor a/k/a/ strict liability.] In this situation, the second agent would, in Nash Equilibrium, take no precaution, and the first agent would adopt a level of precaution in $b_1(0)$. Letting these levels of precaution be x_1^n and x_2^n, and similarly denoting the socially-optimal levels resulting from Coase equilibrium by x_1^c and x_2^c, it is obvious that agent 2 will require a payment of at least $C_2(x_2^c)$ from agent 1. Agent 1, on the other hand, will not be willing to pay more $C_1(x_1^n) - C_1(x_1^c) + \overline{C}(x^n) - \overline{C}(x^c)$: this quantity being the amount he stands to gain from persuading agent 2 to take some precautions. By assumption, total cost to society is less at x^c than at x^n, i.e.,

$$C_1(x_1^c) + C_2(x_2^c) + \overline{C}(x^c) \leq C_1(x_1^n) + C_2(0) + \overline{C}(x^n)$$

from which it follows, given that $C_2(0) = 0$, that there exist payments P satisfying

$$(7) \qquad C_2(x_2^c) \leq P \leq C_1(x_1^n) - C_1(x_1^c) + \overline{C}(x^n) - \overline{C}(x^c);$$

i.e., payments that are acceptable to both parties. Any such payment will do, and the only certain conclusion we can make is that agent 2 is better off under this scheme, which leaves him with a net wealth of $P - C_2(x_2^c) > 0$, than he would be had we reached a Coase equilibrium under the opposite liability rule, $L_2(x) = 1$, where agent 2 is left with a negative net wealth.

There are many ways of refining this result to focus on the distributional implications. For example, we might presume that the market for precautions obeyed the conditions of pure competition, so that each unit of precaution "sold" at a constant price, equal to $C_f'(x_f^c)$. In this case the payment would equal $x_f^c C_f'(x_f^c)$; and this would be feasible only so long as x_f^c exceeded the point of minimum average cost. This second-order inefficiency could be expected to persist, since in this model each agent is a natural monopolist in the production of his own precautionary activity.

Another model might take explicit account of this monopoly power, and would allow each agent to extract the full monopoly profits from his precautionary activity. To represent this solution as a trade optimum, we would need to allow each agent to be a perfectly discriminating monopolist; inasmuch as this results in each agent receiving the whole amount of the inframarginal gains, we shall not further specify it.

There are three other approaches that we shall explore in this section. The first makes use of the Nash model of bargaining with variable threats to decide on an efficient allocation that takes account of the relative bargaining strengths of the various agents. For a complete and thorough discussion of this solution, see Roth.[3] For our purposes, it will suffice to observe that each agent can announce a level of precaution x_f^d that (s)he will adopt in the event of disagreement. The disagreement strategies lead to the <u>disagreement payoff</u>: each agent gets

$$(8) \qquad C_f(x_f^d) + L_f(x^d)\overline{C}(x^d) = H_f(x^d)$$

The set of <u>agreements</u> we shall take to be the set of all efficient and individually-rational levels of precaution, where a level of precaution $\bar{x}$ is <u>individually-rational for f</u> iff

$$(9) \qquad C_f(\bar{x}_f) + L_f(\bar{x})\overline{C}(\bar{x}) \leq \min_{x_f} \max_{x_{(f)}} [C_f(x_f) + L_f(x_f,x_{(f)})\overline{C}(x_f,x_{(f)})] \equiv m_f$$

since agent f can guarantee that he will pay no more than the amount on the RHS of the above equation.

We must also include the possible side-payments between the agents, of course, so that an agreement can be thought of as an n-vector, $a = (a_1,\ldots,a_n)$ with the following properties:

i) $\sum_{f\in N} a_f = \min_x [\sum_{f\in N} C_f(x) + \overline{C}(x)] = C^*$ (efficiency)

ii) for each f, $a_f \geq m_f$ (individual rationality)

Under these circumstances, the Nash Bargaining solution says that, when the players make the threats x^d, they receive the agreement payoff $a(x^d)$ where

$$(10) \qquad a(x^d) \text{ minimizes } (a_1 - H_1(x^d))(a_2 - H_2(x^d))\ldots(a_n - H_n(x^d))$$

In other words, each player f receives

$$(11) \qquad a_f(x^d) = H_f(x^d) + \frac{1}{n}[C^* - \sum_{g \in N} H_g(x^d)]$$

This defines a new payoff function for a noncooperative game, and player f will now select his threat to maximize his agreement payoff, subject to the threats of the other players. This leads to a situation where the players threaten each other with threats x^t satisfying

$$(12) \qquad C_f'(x^t) + \frac{\partial L_f(x^t)}{\partial x_f}\bar{C}(x^t) + L_f(x^t)\frac{\partial \bar{C}(x^t)}{\partial x_f}$$

$$- \frac{1}{n}[\sum_{g \in N}[\frac{\partial L_g(x^t)}{\partial x_f}\bar{C}(x^t) + L_g(x^t)\frac{\partial \bar{C}(x^t)}{\partial x_f}] = 0$$

If we have an agreement, as prescribed above, each agent will actually wind up in a situation where (s)he takes the optimal action, and receives a side-payment from the other players that results in a net cost of $a_f(x^t)$. It is worth noting that as the number of participants goes to infinity, the threats approach precisely those levels of precaution which the agents take at the Nash Equilibrium. Another interesting result is the following: if the liability rules are of the constant cost-sharing variety, the threats by the agents, and therefore the Nash

Equilibrium actions, approach the social optimum. To see this, we first note that the limit of conditions (12) as n tends to infinity is

$$(13) \qquad \text{for all } f,\ C_f'(x^t) + \frac{\partial L_f(x^t)}{\partial x_f}\,\bar{C}(x^t) + L_f(x^t)\frac{\partial \bar{C}(x^t)}{\partial x_f} = 0$$

which is just condition (3) for Nash Equilibrium. On the other hand, inserting the condition that L be a constant cost-sharing rule into (12) gives us

$$(14) \qquad \text{for all } f,\ C_f'(x^t) + \frac{n-1}{n}\cdot\frac{\partial\, \bar{C}(x^t)}{\partial x_f} = 0$$

and the limit of this condition as n tends to infinity is just condition (1) for social optimality.

There are two other possible solutions to the distribution problem that we would like to describe. Both of these stem from a re-casting of the situation as a cooperative game of transferrable utility. In such a game, we are given a set of players and, for each subset of players, or coalition, we are given a number that represents the worth of the coalition; or the total amount of utility or money it can guarantee to its members. This amount that a coalition S can guarantee to its members is denoted v(S). The coalition of the whole, N, has available an amount v(N) and we say that an allocation is the core if, letting $a_1,\ldots,a_n$ denote the allocation, we have

i) $\sum_{f\in N} a_f = v(N)$ (efficiency)

ii) for each $S \subset N$, $\sum_{f\in S} a_f \geq v(S)$ (Coalitional rationality)

The interpretation is that no coalition could do better using its own resources; no coalition could guarantee each of its members a higher payoff.

However, we have not been given a game in this form. Instead, we have started with a strategic game. It is possible to pass from this to a game of transferrable utility that represents the cooperative power of groups of agents. To do this, we must determine what the minimal cost is that a coalition can guarantee to its members. At this point, however, a game-theoretic subtlety arises, since the amount that a coalition can guarantee itself may not be the same as the amount that the other players cannot prevent it from getting. In fact, the general situation is that if a coalition can guarantee itself a certain amount, then it cannot be prevented from getting that amount, but the converse may be false. The reason is that when we use the term "guarantee" we mean that the coalition chooses a certain strategy with the property that it will get at least the specified amount, no matter what the opposition does. In this sense the opposition "moves second." On the other hand, in order to prevent a coalition from getting a certain amount it is necessary for the opposition to "move first." Since the player who moves first is at a disadvantage, since his/their move(s) is known, it follows that anything a coalition can get when it moves first it can certainly get when it moves second.

We shall therefore define two different characteristic functions for our Coase game. The first captures the idea of what a coalition can guarantee itself, and the second the idea of what is cannot be prevented from obtaining.

Definition: Let $S \subset N$; a strategy for coalition S is a strategy for each agent $f \in S$, and is denoted x^S; the strategy adopted by the complementary coalition is denoted $x^{(S)}$. We write $x^S = (x_f^S: \ f \in S)$

$$v_a(S) = \min_{x^S} \max_{x^{(S)}} \sum_{f \in S} C_f(x_f^S) + L_f(x^S, x^{(S)}) \bar{C}(x^S, x^{(S)})$$

$$v_b(S) = \max_{x^{(S)}} \min_{x^S} \sum_{f \in S} C_f(x_f^S) + L_f(x^S, x^{(S)}) \bar{C}(x^S, x^{(S)})$$

The a-core of the Coase game is the set of allocations $a = (a_1, a_2, \ldots, a_n)$ s.t

i) $\sum_{f \in N} a_f = C^*$

ii) for each S, $\sum_{f \in S} a_f \leq v_a(S)$

[Remember that the a_f are costs, and that each agent wants to minimize a_f!] The b-core of the Coase game is defined similarly, except that condition ii) is replaced by:

ii) for each S, $\sum_{f \in S} a_f \leq v_b(S)$

We remark that, for each coalition S, $v_a(S) \geq v_b(S)$, so that the a-core contains the b-core. It follows that showing the non-emptiness of the b-core also suffices to show the nonemptiness of the a-core.

In order to show that the core of our particular game is nonempty, we make use of the following construction [cf e.g., Luce and Raiffa]: Let K be a collection of coalitions, not necessarily disjoint. If $K = [S_1, \ldots, S_k]$, and if each $i \in N$ belongs to at least one member of K, we can define the characteristic vector χ^S of each coalition S to be

that n-vector with 1's in the elements corresponding to members of S and 0's elsewhere. Then, if there exists a collection of numbers $d(S_i)$, one for each S_i in K, which are non-negative and have the property that $\sum_{i=1}^{k} d(S_i)\chi^{S_i} = \chi^N$, then K is said to be a balanced collection and the $d(S_i)$ are balancing weights. A core of a game v is nonempty if and only if, for each balanced collection K and balancing weights d, we have

$$(*) \qquad \sum_{i=1}^{k} d(S_i)v(S_i) \geq v(N)$$

We can now state and prove our main result on the core:

Theorem: Let L be a cost-sharing system of liability rules with the property that, for each agent f, the function $C_f(x_f) + L_f(x)\overline{C}(x)$ is convex in x. Then the b-core of the Coase game is non-empty.

Proof: Let K be a balanced collection of coalitions, and d a set of balancing weights. Moreover, for each $S \in K$, define an n-vector $x(S) = (x_1(S),\ldots,x_n(S)) = (x^S(S),x^{(S)}(S))$ [arranging the players into the coalitions S and N - S] by the requirement that x(S) be any vector of precautions that achieves $v_b(S)$. Define:

$$(15) \qquad \overline{x}_f = \sum_{S\in K} d(S)x_f(S)$$

By definition, $v_b(S) = \sum_{f\in S} C_f(x_f(S)) + L_f(x(S))\overline{C}(x(S))$. By the definition of $v_b(N)$ and the fact that L is a cost-sharing system, we have

$$(16) \qquad v_b(N) \leq \sum_{f\in N} [C_f(\overline{x}_f) + L_f(\overline{x})\overline{C}(\overline{x})] = \sum_{f\in N} C_f(\overline{x}_f) + \overline{C}(\overline{x})$$

To show that the core is nonempty, we need to show

$$v_b(N) \leq \sum_{S \in K} d(S) v_b(S) = \sum_{S \in K} d(S) \left[\sum_{f \in S} [C_f(x_f(S)) + L_f(x(S)) \overline{C}(x(S))] \right]$$

$$\text{RHS} = \sum_{f \in N} \left[\sum_{\substack{S \in K \\ S \ni f}} [d(S)[C_f(x_f(S)) + L_f(x(S)) \overline{C}(x(S))]] \right]$$

and by the assumed convexity, this term is less than or equal to

$$\sum_{f \in N} C_f(\overline{x}_f) + L_f(\overline{x}) \overline{C}(\overline{x}) = \sum_{f \in N} C_f(\overline{x}_f) + \overline{C}(\overline{x})$$

and the theorem is proven. QED

We shall indicate in the next section why we prefer to use the b-core, instead of the more generous a-core. However, we should point out that under our convexity assumptions, which are satisfied for any cost-sharing constant liability rules if $\overline{C}$ and the C_f are convex, the problem raised by Aivazian and Callen cannot arise.

One other solution concept which we should mention is the Shapley value. This is a solution concept for games in characteristic function form which is the unique solution satisfying certain axioms of efficiency and equity. These axioms state that the value allocation should be efficient, that it should give nothing to agents who contribute nothing to the worth of any coalition, that it should depend only on the worth of the players in coalitions and not on outside considerations like names, and finally, that the value in the sum of two games [$w(S) = v(S) + u(S)$] should be the sum of the values in the games. It turns out that the value assigns to each player his or her expected marginal contribution to a random coalition consisting of all players preceding the given player in a random order selected with probability $1/n!$ That is, the Shapley value assigns to player i the amount

$$(17) \qquad \psi v_b(i) = \sum_{S \in i} \frac{(n - s)!(s - 1)!}{n!} [v_b(S) - v_b(S - i)]$$

As A. Roth has pointed out, this represents the expected utility of playing position i in this game, and should be thought of as a benchmark allocation to be used in selecting liability rules on equity grounds. Although it reflects the relative strengths of the players it does so in a different way than the Nash Bargaining solution. In the bargaining solution, the players select optimal threats, where the criterion of optimality is the payoff at the bargain that will be struck eventually between all the players. Here, no coalitions are involved except the individuals and the grand coalition. In the Shapley value, players bargain on the basis of their contribution to the security level of all possible coalitions, but the worth of a coalition is calculated on the basis of defense and not threats. We conclude this section with a simple example in which the Nash Equilibrium is inefficient, both the bargaining solution and the Shapley value are in the core, but the bargaining solution and the value rank the players in reverse order. The liability rule is a modification of Professor Calabrese's idea of the "least-cost-avoider." Players with higher private costs of precautions bear lower levels of liability.

EXAMPLE:

There are three agents, indexed 1,2,3. They face private costs $C_i(x_i) = P_i \cdot x_i$, liability $L_i(x) = \lambda_i$, a cost-sharing constant liability rule, and generate social costs of $\bar{C}(x) = \frac{1}{x_1 x_2 x_3}$.

A. The efficient outcome. Let $Z = \sqrt[4]{P_1 P_2 P_3}$

$x_i^e = \frac{Z}{P_i}$ cost to i = $(1 + \lambda_i)Z$

Social cost = Z

Total cost = C* = 4Z

B. The Nash Equilibrium. Let $w = \sqrt[4]{\lambda_1 \lambda_2 \lambda_3}$

$x_i^n = \frac{\lambda_i}{P_i} \cdot \frac{Z}{w}$ total cost = $\frac{Z}{w}(1 + \frac{Z^2}{w^2})$

social cost = $\frac{Z^3}{w^3}$ cost to i = $\lambda_i \frac{Z}{w}(1 + \frac{Z^2}{w^2})$

C. The b-core. In order to keep the cost to the coalitions of 1 and 2 members finite, we adopt the restriction that some minimal level of precaution is required of all agents. Thus $\forall i \quad x_i \geq \varepsilon > 0$. We then have

$$V_B(i) = \frac{2}{\varepsilon}\sqrt{\lambda_i P_i}$$

$$V_B(i,j) = \frac{3}{\sqrt[3]{\varepsilon}}\sqrt[3]{(\lambda_i + \lambda_j)P_i P_j}$$

$$V_B(1,2,3) = 4Z$$

Then the core is

$\{(a_1,a_2,a_3)$:

1) $\frac{2}{\varepsilon}\sqrt{\lambda_i P_i} \geq a_i \geq \frac{3}{\sqrt[3]{\varepsilon}}\sqrt[3]{(\lambda_j + \lambda_R)P_j P_k}$

2) $\Sigma a_i = 4Z\}$

Before proceeding, let us further specify the example. Let $P_1 = 1$, $P_2 = 2$, $P_3 = 3$ and let us adopt a modified Learned Hand Rule: the lower an agents' cost of precaution, the higher is that agents' liability, $\lambda_i = \frac{a}{P_i}$ so $a = \frac{1}{\Sigma\frac{1}{P_i}}$. Also $\varepsilon = \frac{1}{8}$, $\lambda_1 = .5\overline{45}$, $\lambda_2 = .2\overline{72}$, $\lambda_3 = .1\overline{81}$.

Here we have

$V_B(1) = V_B(2) = V_B(3) = 11.8168$

$V_B(1,2) = 7.0704$; $V_B(1,2) = 7.7820$; $V_B(2,3) = 8.3829$

$V_B(1,2,3) = 6.2603$

core = $\{(a_1,a_2,a_3)$: $a_1 + a_2 + a_3 = 6.2603$

$11.8168 \geq a_1 \geq -2.1226$

$11.8168 \geq a_2 \geq -1.5217$

$11.8168 \geq a_3 \geq -0.8101\}$

Nash equilibrium $Z = 1.5651$ $w = .7212$

$\frac{Z}{w} = 2.1701$ TC = 12.39 (!)

$x_1^N = 1.1837$	payoffs	$L_1^N = 6.7584$
$x_2^N = 0.2959$		$L_2^N = 3.3792$
$x_3^N = 0.1315$		$L_3^N = 2.2528$

Efficient Outcome

	Payoffs (uncompensated)
$x_1^e = 1.5651$	$L_1^e = 2.4188$
$x_2^e = 0.7826$	$L_2^e = 1.9919$
$x_3^e = 0.5217$	$L_3^e = 1.8497$

Nash Bargaining solution

	Coop. Payoffs
threats: $x_1^t = 1.4142$	$a_1 = 2.2459$
$x_2^t = 0.7071$	$a_2 = 2.0413$
$x_3^t = 0.4714$	$a_3 = 1.9373$

Shapley Value

$$\psi V_B(1) = 1.7675$$

$$\psi V_B(2) = 2.0690$$

$$\psi V_B(3) = 2.4244$$

III. A Decentralized Rationale for the Core

In this section, we take a different approach to the distributional problem. With the core, distribution was decided during a once-and-for-all bargaining session, with the bargain being enforced by an outside agency. There, the relevant question was whether any coalition could do better using its own resources. On the other hand, as Professor Coase[2] points out, it is likely that such agreements will be secured by a series of contracts, including penalty clauses. These arrangements have the effect of changing the incentives acting on the parties in such a way that their long-run interest is to adhere to the terms of the contract without the need to invoke an outside agency. What is important is the punishment that the defectors will suffer at the hands of the remaining parties. It turns out that the set of agreements that can be supported by such contracts is precisely the b-core.

In such an arrangement, time is involved explicitly, and future behavior is made conditional on past performance, so that such social institutions as threats and promises form part of the agreement. Perhaps

the simplest model of this situation is that of the supergame, in which the Coase game described in the previous section is played repeatedly, and where the actions taken at the t^{th} play of the game are allowed to depend on the history of play up to time t.

Formally, if the actions available to each player in each play of the game are X_f, and if $X = \underset{f \in N}{\times} X_f$ denotes the set of outcomes of each play, a history of length t is just an element of X x ... x X (t times), which we write X^t. A strategy for player f is a collection $s_f = (s_f^1,\ldots,s_f^t,\ldots)$ of functions where:

$$s_f^1 \in X_f, \text{ and}$$

$$s_f^t : X^{t-1} \to X_f \qquad \text{for all } t > 1$$

If each player $f \in N$ selects such a strategy, we can calculate the sequence of outcomes $x(s) + [x^1(s),\ldots,x^t(s)\ldots]$, where $x^t(s)$ is the play of the game at date t that results from the strategy choice $s = (s_1,\ldots,s_n)$:

$$x^1(s) = r^1(s) = (s_1^1,\ldots,s_n^1)$$

$$\vdots$$

$$x^t(s) = [s_1^t(r^{t-1}(s)),\ldots,s_n^t(r^{t-1}(s))]$$

$$r^t(s) = [r^{t-1}(s), x^t(s)]$$

So the x(s) are the outcomes, and the r(s) are the partial histories. Another result of the strategy choice is an infinite stream of costs $[p_1^1,\ldots,p_n^1,p_1^2,\ldots]$ where

$$p_f^t(s) = C_f(x_f^t(s)) + L_f(x^t(s))\bar{C}(x^t(s))$$

We shall start out by assuming that each agent wishes to minimize the limiting average of these costs:

$$H_f(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t\leq T} p_f^t(s)$$

assuming that it exists. We shall sidestep the existence question, since it will not affect our results--the interested reader is referred to Aumann[4] for details. Later on, we shall replace this evaluation of the cost stream by the discounted present value at discount rate δ:

$$H_f^\delta(s) = (1-\delta) \sum_{t=1}^{\infty} \delta^{t-1} p_f^t(s)$$

which is more intuitive.

Without going into great detail, we shall argue that any supergame strategy combination can be interpreted as a contract; that a certain type of equilibrium, called a strong equilibrium, describes a stable contract; and finally that the set of limiting average costs to strong equilibria of the supergame coincides with the b-core of the one-shot Coase game.

Any combination of supergame strategies may be described in two parts. The first part, represented by x(s), is the "specified behavior": what happens when things go according to plan. The second part, which includes all the behavior called for by s after histories $r^t \neq r^t(s)$. that are not expected, can be termed a penalty clause, it specifies the

behavior that parties will adopt in the event of a deviation from the specified behavior, or a breach of the contract.

Recall that a Nash Equilibrium was defined to be a situation from which no party could profitably and unilaterally defect. Of course, this is not sufficient for our purposes in a world in which collusion, cartelization and conspiracy are possible, especially in the repeated-game context where coordination becomes relatively easy. To account for these features, we shall describe a situation from which no coalition, or subset of players, can profitably defect, given that the remaining players adhere to their original strategies.

Formally, let us say that s* is a strong equilibrium of the supergame if, for any coalition $F \subset N$, letting s^F denote the strategies of members of F, we have

$$(18) \qquad \sum_{f \in F} H_f(s^*) \leq \sum_{f \in F} H_f(s^F, s^{*(F)}) \qquad \text{for all } s^F$$

Of course, we could have defined strong equilibria for the one-shot game. However, since one of the coalitions F is the "coalition of the whole", N, every strong equilibrium is a fortiriori Pareto Optimal, or efficient, as well as being a Nash equilibrium. Since it is usually the case that the Nash equilibria of the Coase game are inefficient, it follows that there are usually no strong equilibria.

On the other hand, the contract represented by a strong equilibrium of the supergame is stable against any defection, including unanimous rejection by all the parties. To find out which outcomes of the one-shot game can be supported by such stable contracts, it suffices to look at strategies of a very simple form called "grim" strategies. A grim

strategy has two major parts: the first is a cooperative sequence, specifying an action for each player on each day. The limiting average costs obtainable in this way include every convex combination of pure-strategy costs in the one-shot game. The other part of a grim strategy is a punishment sequence. If a coalition defects from the cooperative sequence at any stage, this part of the strategy calls for the strongest possible punishment to be inflicted on them forever. Since the punishment must be specified in the strategy, it follows that the defecting coalition can choose a defence in light of the opposition's punishment. This means that the defectors cannot be forced to pay more than

$$(19) \qquad M_F = \max_{x^{(F)}} \min_{x^F} \sum_{f\in F} [C_f(x_f^F) + L_f(x^F,x^{(F)})\overline{C}(x^F,x^{(F)})]$$

at any stage of the game. It therefore follows that the strongest punishment that can be inflicted on a defecting coalition F is to hold it to its "maxmin" level M_F forever. This in turn means that F will only defect if its total average cost along the cooperative sequence exceeds the amount M_F. The reason for this is that any savings earned at the beginning of the defection vanish during the long-term punishment. The actual average payoff to F will then consist of a certain amount C(F) paid out before the punishment begins, and an unending sequence of amounts at least as large as M_F afterwards. The long-term average payoff is therefore at least

$$\lim_{T\to\infty} \frac{C(F)}{T} + M_F = M_F$$

We have therefore demonstrated what we set out to show:

Theorem 3.1: $a \in R^n$ can be achieved as the limiting average cost to a strong equilibrium (stable contract) in the supergame if and only if, for each $F \subset N$

$$(20) \qquad \sum_{f \in F} a_f \leq \max_{x^{(F)}} \min_{x^F} \sum_{f \in F} [C_f(x_f^F) + L_f(x^F, x^{(F)}) \overline{C}(x^F, x^{(F)})$$

which is exactly equivalent to the condition "a is in the b-core of the one-shot Coase game."

For a more formal proof of this "Folk Theorem", the reader is referred to Aumann[4] or Cave.[5] The result relies heavily on the fact that no savings accrued during a finite period of time can affect the limiting average cost. This seems inherently unreasonable, and we should expect the players to discount the future. This discounting weakens the effect of the grim punishments, and consequently diminishes the set of outcomes that can be supported by stable contracts. To capture the effects of this myopia more precisely, we shall observe that the reasoning used above still works if the players use the discounted evaluation relation $H_f^\delta(s)$ defined above. However, in this case, the cost savings due to the defection in the first period become important, and the theorem must be modified to take account of this.

This is slightly complicated by the existence of sidepayments between the players. The simplest assumption is that, while everyone takes the optimal actions everywhere along the cooperative sequence, the sidepayments cease immediately whenever there is any irregularity

in the players' actions. In this case, the defecting coalition will act to minimize its cost during the period of defection, given that the other players continue to take the efficient actions x^e:

$$\min_{x^F} \sum_{f\in F} [C_f(x_f^F) + L_f(x^F, x^{e(F)})\overline{C}(x^F, x^{e(F)})]$$

call this "best defection" cost BD_F. The scenario is now as follows: because of the discounting, the coalition may as well defect now as at any future time. If they do not defect, they pay a discounted present value of $\sum_{f\in F} a_f$ (notice that we have normalized by multiplying by $(1-\delta)$). On the other hand, if they defect today, and receive the grim punishment in the future, they pay a cost of BD_F today, and M_F in all subsequent periods. If we call the set of outcomes that can be sustained as the result of stable contracts in the discounted supergame the <u>δ-core</u>, we have the following result:

<u>Theorem 3.2</u>: $a \in R^n$ is in the δ-core if and only if, for each coalition $F \subset N$,

$$(21) \qquad \sum_{f\in F} a_f \leq (1-\delta)BD_F + \delta M_F$$

It will be noticed that we get the previous result by letting the discount rate go to 1, and that we get the definition of strong equilibrium for the one-shot game by letting the discount rate go to 0. For a formal proof of this proposition, the reader is referred to Cave.[5]

At this point, it is probably instructive to return to the example used in the previous section, to see what effect discounting has on the distributions we might expect in the Coase game.

Example: Recall that there are three agents, and that agent f is characterized by:

$$C_f(x_f) = p_f x_f$$

$$L_f(x_f) = \left[\sum_{g \in N} \frac{1}{p_g} \right]^{-1}$$

$$\bar{C}(x_1, x_2, x_3) = \frac{1}{x_1 x_2 x_3}$$

Moreover, we define $Z = [p_1 p_2 p_3]^{1/4}$, $p_1 = 1$, $p_2 = 2$, $p_3 = 3$. Finally, we have a minimum precaution restriction, $x_f \geq 1/8$ for each f. To begin with, the efficient actions are given by

$$x_1^e = 1.5651 \qquad x_2^e = 0.7826 \qquad x_3^e = 0.5217$$

and result in a total cost to society of 6.2603. If we denote the "best defection" strategies for members of a coalition F by y_f^F, $f \in F$, we find that

$$y_1^1 = 1.559 \qquad y_1^{1,2} = 1.4638 \qquad y_1^{1,3} = 1.4075$$

$$y_2^2 = 0.4087 \qquad y_2^{1,2} = 0.7319 \qquad y_2^{2,3} = 0.6017$$

$$y_3^3 = 0.2225 \qquad y_3^{1,3} = 0.4692 \qquad y_3^{2,3} = 0.4011$$

(and $y_f^{1,2,3} = y_f^e$ for each f). This allows us to calculate the best defection costs BD_F for each coalition F. In the table below, we have listed BD_F and M_F.

F	BD_F	M_F
1	2.3117	11.8168
2	1.6347	11.8118
3	1.3347	11.8168
1,2	4.3914	7.0704
1,3	4.2223	7.7820
2,3	3.6101	8.3829
1,2,3	6.2603	6.2603

In describing the conditions for the δ-core, we shall rearrange the condition given in the theorem to:

$$(22) \qquad \sum_{f\in F} a_f \leq BD_F + \delta[M_F - BD_F]$$

Moreover, we can write all seven conditions for the core, one for each of the seven nonempty coalitions, in a simplified form by observing that if (22) is satisfied for both F and N-F, since $\sum_{f\in F} a_f + \sum_{f\in N-F} a_f = M_N = 6.2603$, we may write

$$(23) \qquad M_N - BD_{N-F} - \delta[M_{N-F} - BD_{N-F}] \leq \sum_{f\in F} a_f \leq BD_F + \delta[M_F - BD_F]$$

It follows that the conditions for the δ-core of our three person game can be written as conditions in the individual costs a_1, a_2, and a_3 as follows:

(24) i) $a_1 + a_2 + a_3 = 6.2603$

ii) $2.6502 - (4.7728)\delta \leq a_1 \leq 2.3117 + (9.5051)\delta$

iii) $2.038 - (3.5597)\delta \leq a_2 \leq 1.6347 + (10.1821)\delta$

iv) $1.8689 - (2.6790)\delta \leq a_3 \leq 1.3347 + (10.4821)\delta$

It can be readily seen that for some values of the discount rate the δ-core is empty; in fact, for any value of the discount rate less than 0.0406 this will be the case.

There are several further questions that we can answer with these conditions. These are: what is the b-core; what are the first allocations to appear in the δ-core as the discount rate rises, and for what values of δ is the "no-transfer point" where each party takes the correct action but no sidepayments are made a part of the δ-core?

To find the b-core we substitute $\delta = 1$ into equation (24) giving the conditions obtained in Section II.

To find the first core points we insert the crucial value $\delta = 0.0406$, giving

(25) $2.4564 \leq a_1 \leq 2.6976$

$1.8935 \leq a_2 \leq 2.0481$

$1.760 = a_3$

We can write this more succinctly as

$\{(a_1, 4.5003-a_1, 1760): 2.4564 \leq a_1 \leq 2.6068\}$ making use of the fact

that the a_f sum to 6.2603. As to the final question: we can determine that for the no-transfer point to be in the δ-core we need

$$(26) \qquad 2.6502 - (4.7728)\delta \leq 2.4118 \leq 2.3117 + (9.5051)\delta$$

$$2.038 - (3.5597)\delta \leq 1.9919 \leq 1.6347 + (10.1821)\delta$$

$$1.8689 - (2.679)\delta \leq 1.8497 \leq 1.3347 + (10.4821)\delta$$

$$.04995 \leq \delta$$

$$.01532 \leq \delta$$

$$.01295 \leq \delta$$

$$.03508 \leq \delta$$

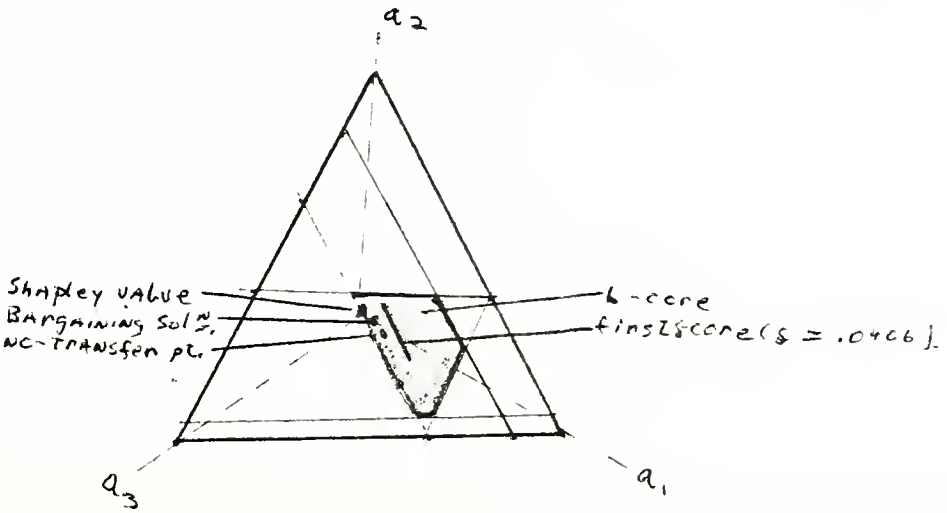
$$.00717 \leq \delta$$

$$.04913 \leq \delta$$

$$\delta \geq \underline{.04913} > .0406$$

What of the other proposed efficient solutions: by analagous calculations, we can determine that for the Nash bargaining solution to be in the δ-core we must have $\delta \geq \max\{.08471, .03993, .05749\} = \underline{.08471}$ and for the Shapley value: $\delta \geq \max\{.1849, .04265, .10393\} = \underline{.1849}$.

In the following figure we have illustrated the b-core, the δ-core for $\delta = 0.0406$, the no-transfer point, the value and the bargaining solution.



strategy s_f^t had to specify an action for any member of X^t. A subgame is a supergame that starts on date t, following some arbitrary history of length t-1, and a strong perfect equilibrium of the supergame is a strategy combination, s, that specifies equilibrium behavior in every subgame. In terms of the previous discussion, this means that the defecting coalition will show that the specified punishment is, in fact, the best thing for the remainder of the players to do.

Definition: s is a strong perfect equilibrium iff, for every t, and every partial history $x^t \in X^t$, the strategy $\bar{s}(\quad :x^t)$ defined by:

$$\bar{s}^{t'}(y^{t'-1}:x^t) = s^{t+t'}(x^t,y^{t'-1})$$

is a strong equilibrium in the original supergame.

It is clear from this definition that the grim punishments will not work, in general, so that one might anticipate that not all strong equilibrium outcomes could be supported by credible threats. However, in the undiscounted game, this anticipation is confounded.

Theorem 4.1: the set of outcomes $a \in R^n$ that can be supported by strong perfect supergame equilibria (credible and stable contracts) coincides with the set of outcomes that can be supported by strong equilibria; i.e., it is the b-core of the one-shot game.

For details of the proof, the reader is referred to Cave.[5] However, the idea of the proof is simple enough: to support a given outcome of a strong equilibrium by credible threats, we recall the following apparatus from the grim punishment: to begin with, we have the cooperative sequence $(x^{c1},\ldots,x^{ct},\ldots)$ which gives the desired

IV. The Credibility of Strong Supergame Equilibria

In addition to the "never-never-effect" discussed above, there is another feature of the grim strategies which may make them seem unreasonable. For coalitions which receive nearly their security level payoffs, the only effective punishments are those which actually hold them to this level for a long period of time. Consider a world in which the liability rules are constant; in such a world, the only way to punish a coalition F is to maximize the social cost in each period. In addition to increasing F's cost, this increases the cost to the punishing coalition. This increased cost may be so high that F may not believe that N-F would actually carry out the planned punishment, especially if the alternative is to continue receiving the payoff specified by the cooperative sequence. This credibility problem can be avoided if we require that the equilibrium strategies specify optimal behavior in every eventuality, and not just in those situations which can arise by individual strategy variations.

In other words, when we defined a strong equilibrium, we asked each coalition, taking the actions of the complementary group as fixed, to examine the outcomes it could achieve by varying its own strategy. The requirement was that the stated behavior give the best of these outcomes. However, the behavior in situations that could not arise according to the strategies used by the other players was completely arbitrary. We shall now strengthen this by defining the concept of a subgame. Recall that a partial history of length t was defined to be a member of X^t. We distinguished the particular member of X^t that resulted from the use of strategies s by $x^t(s)$, but each t^{th} period

cost $a \in R^n$ as a limiting average. Moreover, for each coalition F, there is a <u>punishment</u> $x^F_{(F)}$ to be employed against F by the members of N-F, which has the property that the best F can do against $x^F_{(F)}$ is to get M_F in any period where $X^F_{(F)}$ is being employed against them. Now, let us define a sequence of strictly positive real numbers $e^1,\ldots,e^t,\ldots$ with the property that $\lim_{t\to\infty} e^t = 0$. For any sequence $(x^1,\ldots,x^t)$ of finite length, we may define the cumulative average cost to player f for each f, and also for each coalition F. Thus, we define

$$C^*(x,F,t) = \sum_{t'\leq t} \sum_{f\in F} \frac{1}{t} C_f(x_f^{t'}) + L_f(x^{t'})\overline{C}(x^{t'})$$

The cumulative average costs will be used to measure the amount of punishment each player or each coalition has suffered. The strategy will specify that a coalition that defects at period t will be punished until the first time t' when

$$(26) \qquad C^*(x,F,t') \geq M_F - e^t$$

This will happen in finite time, since the e^t are all positive numbers. At this point, play will return to the cooperative sequence. If all players adhere to this plan, the coalition F will be left with a limiting average cost which agrees with their cooperative-sequence cost (while they are left with some savings when their punishment ends, these savings disappear in the limit). However, we are still faced with the original problem: why should play return to the original sequence? It is clear that no player is affected by the cost of punishment, but if a coalition defects forever, the strategy will inflict punishment forever. What we need is a mechanism that ensures that players who do

not carry out their specified punishments are themselves punished. To do this, we must define the idea of a "last defection". Suppose that we are given a collection $s = (s_1,\ldots,s_n)$ of strategies and an arbitrary history $y = (y^1,\ldots,y^t)$ of finite length. We can identify the <u>last defectors</u> LD(s,y) and the time $t^*(s,y)$ of their defection as follows:

$$(27)\qquad t^*(s,y) = \begin{cases} \max\{t^*:\ y^{t^*} \neq s^{t^*}(y^1,\ldots,y^{t^*-1})\} \text{ if it exists} \\ t+1 \text{ otherwise} \end{cases}$$

$$(28)\qquad LD(s,y) = \begin{cases} \{f \in N:\ y_f^{t^*(s,y)} \neq s_f^{t^*(s,y)}(y^1,\ldots,y^{t^*(s,y)-1})\} \\ \qquad\qquad \text{if } t^*(s,y) \leq t \\ \emptyset \text{ otherwise} \end{cases}$$

Finally, we must restrict our attention to those last defectors who have not paid their debt to society as of the current date (t+1):

$$(29)\ D(s,y) = \begin{cases} LD(s,y) \text{ if } LD(s,y) \neq \emptyset \text{ and } \sum_{f\in LD(s,y)} \sum_{\tau\leq t} C_f(y_f^\tau) + L_f(y^\tau)\overline{C}(y^\tau) \\ \qquad < M_{LD(s,y)} - e^{t^*(s,y)} \\ \emptyset \text{ if not} \end{cases}$$

Recalling that the punishments x^F to be used against a coalition F are those defined by equation (20), interpreted as x_f^F for all $f \in N$, we can now write the strategy that gives the cooperative sequence $x^{c1},\ldots,x^{ct}\ldots$ as a strong perfect equilibrium outcome:

$$s_f^1 = x_f^{c1} \text{ for all } f, \text{ and for any partial history } y = (y^1,\ldots,y^{t-1})$$

(30)

$$s_f^t(y) = \begin{cases} x_f^{D(s,y)} & \text{if } D(s,y) \neq \emptyset \\ \\ x^{ct} & \text{otherwise} \end{cases}$$

It is now clear why this is a strong perfect equilibrium. Consider any partial history. If we are on the cooperative sequence, any coalition can choose to defect. However, if they defect for a finite number of periods, they return to the cooperative sequence after a finite number of periods, and make no savings in the limit. If, on the other hand, they defect forever, they will be punished forever, and will be forced to pay at least M_F in average. Moreover, they will believe in the punishments, since any coalition that fails to carry out its share of a specified punishment is itself punished. Finally, we observe that even if punishment costs the punishers more than M_{N-F}, they still would rather carry out the punishment, which they expect to end in finite time. In other words, by carrying out the punishment they guarantee themselves the cooperative payoff in the limit, while by defection (i.e., failing to punish) they ensure that they will be forced to pay M_{N-F}.

In the undiscounted game, we have already seen that the set of outcomes that can be supported by stable and credible contracts is the b-core. Is there an analogous result for the discounted game, saying that the set of outcomes that can be supported by credible and stable contracts coincides with the δ-core? Unfortunately, there is not, except in a few simple cases. For details of these cases, the reader is

referred to Cave.[5] The basic idea is that whenever for each coalition F there exists a strong equilibrium of the supergame giving the members of F a discounted present cost of M_F, we can construct a "grim" strategy that switches to the indicated "punishment equilibrium" whenever F defects. However, in general this will not be the case, as we can show with a simple example.

Example: There are two participants. Player 1 can take a level of precaution x_1, which costs him nothing. If he does so, there is a social cost of $(1-x_1)\overline{C}$, where $\overline{C}$ is a large number. Player 2 cannot take any precautions, but can compensate player 1 by paying him an amount x_2 from her initial wealth of 1. Player 1 pays a constant share, L, of the social cost, and player 2 pays the balance. It follows that the payoffs (net wealths) of the two players given the moves x_1, x_2 are:

$$P_1(x_1,x_2) = x_2 - L(1-x_1)\overline{C}$$

$$P_2(x_1,x_2) = 1 - x_2 - (1-L)(1-x_1)\overline{C}$$

In the one-shot game, there is only one equilibrium: it involves player 1 setting $x_1 = 1$, and player 2 setting $x_2 = 0$. Since it is Pareto optimal, and there are only two players, it is also the only strong equilibrium of the one-shot game. In the undiscounted supergame, the strong equilibrium outcomes are exactly those allocations (a_1,a_2) of net wealth with the following properties:

i) $a_1 + a_2 = 1$ (Pareto Optimality)

ii) $a_1 \geq 0$ (individual rationality for player 1)

iii) $a_2 \geq 1 - (1-L)\overline{C}$ (individual rationality for player 2)

It appears then, that player 1 can use the threat of taking no precaution to extract some payment from player 2. Moreover, by Theorem 4.1 we know that this threat can be made credibly, so that player 1 could wind up with a net profit of as much as $(1-L)\overline{C}$.

Now suppose that we are playing the discounted game, and we wish to see whether we can get the result $x_1 = 1$, $x_2 = 1$ with a stable, and credible contract. To begin with, in order to get this result, which gives player 1 a perpetual payoff of 1 and player 2 a perpetual payoff of 0, from a stable contract, it must be that neither individual player can profit by unilateral defection. Player 1 will not defect, since there is no way he can hope to profit by increasing the social loss. Player 2, on the other hand, can defect in any period for an immediate payoff of 1. Under the grim strategies, this defection will mean that player 2 will get at most $1 - (1-L)\overline{C}$ in all subsequent periods. If both players are using the same discount rate d, this means that player 2 will defect unless

$$0 \geq (1-d) \cdot 1 + d \cdot [1-(1-L)\overline{C}], \text{ i.e., unless}$$

$$(31) \qquad (1-L)\overline{C} \geq \frac{1}{d}$$

Now, suppose that player 1 is to punish player 2 for some defection, by playing a sequence $(x_1^1,\ldots,x_1^t,\ldots)$. The total punishment that is inflicted on player 2 is

$$(32) \qquad \sum_{t=1} d^{t-1}[(1-L)(1-x_1^t)\overline{C}]$$

while the cost to player 1 is

$$(33) \qquad \sum_{t=1} d^{t-1}[(L)(1-x_1^t)\bar{C}]$$

As the ratio of these two quantities is a constant $(\frac{1-L}{L})$, player 1 might just as well punish sooner rather than later, in terms of cost-effectiveness. Therefore, we can limit attention to strategies where player 1 reacts immediately to any defection with some x_1 that sufficiently punishes player 2:

$$(34) \qquad (1-x_1)(1-L)\bar{C} \geq \frac{1}{d}$$

The longer 1 holds off on punishing 2, the cheaper punishment becomes, so the extremes are: punish immediately with x_1 and return to the cooperative sequence for an expected payoff of $\frac{d}{1-d} - (1-x_1)L\bar{C}$, or hold off forever, which gives 1 a payoff of 0. The condition for player 1 to be willing to actually punish player 2 is then

$$0 \leq \frac{d}{1-d} - L(1-x_1)\bar{C} \quad \text{or}$$

$$(35) \qquad \frac{d}{(1-d)L\bar{C}} \geq x-x_1$$

Combining this with (34) gives the following condition for (1,0) to be a payoff sustainable by stable, credible contracts in the discounted supergame:

$$(36) \qquad \frac{d}{(1-d)L\bar{C}} \geq (1-x_1) \geq \frac{1}{d(1-L)\bar{C}}$$

From this equation it is clear that, if

$$(37) \qquad \frac{d^2}{1-d} < \frac{L}{1-L}$$

There is no way to get this outcome via credible stable contracts. In fact, we can use this analysis to calculate precisely the set of distributional outcomes supportable by stable credible contracts for this case. If we consider any outcome (a, 1-a) satisfying the conditions for strong equilibrium in the undiscounted game, we must first see whether they are in strong equilibrium in the discounted game. As before, it is only player 2 who has an incentive to defect. Optimal defection offers her a payoff of $(1-d) \cdot 1 + d \cdot [1-(1-L)\overline{C}]$, while not defecting pays her 1-a. Thus, there is a strong equilibrium of the discounted game with the result (a,1-a) in each period iff

$$(38) \qquad d(1-L)\overline{C} \geq a$$

As before, we may as well assume that player 1 responds immediately to any defection on the part of player 2, employing some x_1 with the property that:

$$(39) \qquad (1-x_1)(1-L)\overline{C} \geq a/d$$

As before, the condition for player 1 to be willing to carry out this threat is given by inequality (35), and we may combine the two to obtain the condition for (a,1-a) to be sustainable by stable, credible contracts:

$$(40) \qquad \frac{d}{(1-d)L\overline{C}} \geq \frac{a}{d(1-L)\overline{C}}$$

from which we obtain the original condition by setting a = 1. In fact, the condition on a can be seen to be independent of $\overline{C}$; it is just:

$$(41) \qquad \frac{(1-L)d^2}{L(1-d)} \geq a \geq 0$$

and this is a complete characterization of the strong perfect equilibrium outcomes, when combined with condition (38).

This result restores some of our intuition about the distribution of the inframarginal gains. For example, credibility limits player 1's payoff most when the discount rate is 1/2 and when the liability rates are equal. If player 1 bears all the liability, a = 0, while if player 2 bears all the liability, 1 can extract up to $d\bar{C}$. In order for credibility to limit the set of attainable outcomes, it must be the case that $(1-d)L\bar{C} \geq d$, so the discount rate must be smaller than $\frac{L\bar{C}}{1+L\bar{C}}$. In other words, this secondary consideration becomes active as the players become more myopic or as the first player's liability becomes smaller.

References

[1]V. A. Aivazian and J. L. Callen, The Coase Theorem and the Empty Core 24 J. Law & Econ. 1 (1981)

[2]R. H. Coase, The Coase Theorem and the Empty Core: A Comment 24 J. Law & Econ. 1 (1981)

[3]A. E. Roth, Axiomatic Models of Bargaining New York, Springer-Verlag 1979

[4]c. f. e.g. R. J. Aumann, Acceptable Points in General Cooperative n-Person Games, in Contributions to the Theory of Games, v. IV (Ann. Math. Stud. 40) A. W. Tucker and R. D. Luce (ed.) Princeton: Princeton University Press (1959)

[5]J. Cave, Equilibrium Behavior in Infinitely-repeated Games Ph.D. Disseration, Stanford Univ. (1980)

[6]R. H. Coase, The Problem of Social Cost 3 J. Law & Econ. 1 (1960)